DOCUMENT RESUME

ED 400 267 TM 025 371

AUTHOR Wise, Steven L.
TITLE A Critical Analysis of the Arguments for and against Item Review in Computerized Adaptive Testing.
PUB DATE Apr 96
NOTE 26p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Gains; *Adaptive Testing; *Computer Assisted Testing; Error Correction; Guessing (Tests); *Responses; *Review (Reexamination); Scores; Test Construction; Testing Problems; *Test Items; Test Results; Test Wiseness; Timed Tests
IDENTIFIERS *Answer Changing (Tests); Item Dependence; Stakeholders

ABSTRACT

In recent years, a controversy has arisen about the advisability of allowing examinees to review their test items and possibly change answers. Arguments for and against allowing item review are discussed, and issues that a test designer should consider when designing a Computerized Adaptive Test (CAT) are identified. Most CATs do not allow examinees the opportunity to review their items. The reasons advanced for this position include: (1) the possibility of item dependence that might affect another answer; (2) a decrease in testing efficiency; (3) opening the test results to effects of test-taking strategies; (4) an increase in testing time; and (5) complications in test development. Arguments in favor of allowing review focus on legitimate score gain possibilities. The first usually advanced is that examinees prefer to be able to review, and the second is that review yields legitimately improved scores. Consideration of arguments for and against item review is complicated by the presence of multiple stakeholders in the measurement process. The question of allowing item review is one without a clear answer, but the interests of test takers and test givers should be protected, perhaps by the development of new types of CAT. (Contains 1 table and 23 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
STEVEN L. WISE

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 400 267

A Critical Analysis of the Arguments For and Against Item Review in Computerized Adaptive Testing

Steven L. Wise

University of Nebraska-Lincoln

Paper presented at the annual meeting of the National Council on Measurement in Education, New York, April, 1996

TM025371

BEST COPY AVAILABLE

A Critical Analysis of the Arguments For and Against Item Review in Computerized Adaptive Testing

The arrival of any new technology is inevitably accompanied by questions regarding its proper use. One such technology in the field of educational measurement, computer-based testing, is making practicable an increasing number of measurement possibilities, such as adaptive testing and new types of test items. Computer-based testing has also provided the possibility of greater control over how examinees take tests; we can now consider testing practices that were previously difficult to implement, such as denying item review or imposing response time limits for each item.

The most visible application of the computer thus far in the field of educational measurement has been the adaptive test. A computerized adaptive test (CAT) administers items[1] whose difficulties are tailored to an examinee's proficiency level, based on his/her item responses to previously administered items. Because of this matching of items to examinees, a CAT is more efficient than a conventional test—typically requiring about half as many items to attain an equivalent precision of proficiency estimation. As a result, CATs are being used in an increasing number of testing programs.

In recent years, a controversy has emerged regarding the advisability of allowing examinees to review their test items, and possibly change answers, when taking a CAT. Some researchers and practitioners feel that item review is illogical and unwarranted in the context of a CAT, while others feel that denying examinees the opportunity to review items is unacceptable and affects the quality of measurement. The purposes of this paper are to (a) discuss the points of argument for and against item review, (b) provide a critical analysis of the arguments in light of relevant empirical research, and (c) identify issues that

should be considered by practitioners when deciding whether or not to provide item review when designing a CAT.

Terminology and Assumptions

Prior to delineating the arguments, pro and con, regarding item review, it will be useful to clarify several terms that I will use in this discussion. In addition, I will note some important assumptions upon which my analysis and recommendations are based.

Review Versus Preview of Items

It is useful to distinguish between the review and preview of test items. The term *review* refers to the examinee behaviors of evaluating answers previously given to items, with the possibility of changing answers for one or more items. In contrast, the term *preview* refers to the examinee behaviors of either (a) browsing through test items before providing answers or (b) inspecting items and choosing to defer answering until later in the test. In the present discussion, only item review is considered.

Goal of Measurement

I am assuming that, when a CAT is used, the goal of measurement is to estimate the proportion of items from a domain of interest for which the examinee knew the correct answer prior to the administration of the test. Hence, the examinee's domain score—the level of an examinee's proficiency— is of primary interest, which implies a domain-referenced perspective on the measurement process rather than a norm-referenced perspective. In item response theory (IRT), which is used to estimate examinee proficiency and identify items to be administered in virtually all operational CATs, proficiency estimates can be readily translated into domain scores. This isomorphism between proficiency estimates and domain scores implies that, given a

representative item pool, a proficiency estimate should indicate an examinee's degree of mastery of the domain of interest.

Item Review and Answer Changing

I am also assuming that a key reason that examinees review test items is to have an opportunity to change answers, if they so desire. That is, relatively few examinees would devote time and effort to item review if no answer changing would be permitted. This assertion that examinees review items primarily for *strategic* purposes does not, however, preclude other potential reasons for item review. For example, highly test anxious examinees might engage in item review for a variety of *affective* reasons (e.g., to gain a perception of control over a stressful testing situation). Although this discussion will be premised on answer changing being the primary reason for item review, empirical research is needed to provide a fuller understanding of why examinees are motivated to review items.

When an examinee reviews an answer to a test item, he/she may choose to change that answer. This action has one of three consequences: (a) a previously incorrect answer is changed to the correct answer, which increases the examinee's test score, (b) a correct answer is changed to an incorrect answer, which decreases the examinee's test score, or (c) an incorrect answer is changed to a different incorrect answer, which does not affect the examinee's test score.

Since the 1920's, there have been more than 60 studies of the effects of answer changing on test performance. Benjamin, Cavell, and Shallenberger (1984) reviewed 33 of these studies, reporting that the median percentage of examinees who changed one or more answers was 84%. The median percentage of answers changed, however, was relatively low (3%). Waddell and Blankenship (1995), who performed a meta-analysis of 61 studies, reported

similar findings; across all studies, 83% of the examinees changed at least one answer, and the mean percentage of answers changed was 5%.

Despite the persistent belief held by many test givers and examinees that answer changing is likely to lower test scores, research has shown that answer changing tends to improve test performance. Benjamin et al. (1984) found that the results of every study in their review were consistent with the conclusion that (a) the majority of answer changes are from incorrect to correct and (b) most examinees who change their answers improve their test scores. These findings are summarized in Table 1, along with those of Waddell and Blankenship (1995). The general effect of answer changing on test performance is therefore quite clear—answer changing is far more likely to result in a score gain than a score loss. For this reason, though some examinees (roughly 15%) experience lower test scores, the general effect of answer changing will be characterized in the present discussion as a score *gain*. It is the interpretation of score gains, moreover, that largely distinguishes the arguments in favor of and against item review.

Legitimate Versus Illegitimate Score Gains

In assessing the psychometric impact of score gains that result from answer changing, it is important to consider the various reasons why examinees change answers from incorrect to correct. I will classify score gains that result from each reason as either *legitimate* or *illegitimate*. Whenever an examinee, at the beginning of the test, possessed the knowledge required to correctly answer a given item, any answer change that results in an incorrect answer being changed to correct is deemed a legitimate score gain, because the final item score more accurately reflects the examinee's state of knowledge regarding the test item. For example, consider an examinee who knew the answer to an item but mistakenly

entered the wrong answer. If, during item review, he/she realized the mistake and then provided the correct answer, the resulting score gain is legitimate.

Illegitimate score gains reflect instances in which an answer was changed from incorrect to correct when the examinee did not possess the knowledge required to correctly answer the item. For example, suppose that the answer to one item was revealed by the information in a second item. If an examinee lacking knowledge initially entered an incorrect answer to the first item, recognized the clue contained in the second item, and consequently provided the correct answer to the first item during review, the resulting score gain is illegitimate because the final item score would not accurately reflect the examinee's initial state of knowledge regarding the first item.

Arguments Against Providing Review

Currently, most of the testing programs that use CATs do not provide examinees an opportunity to review their items. A variety of reasons have been offered to justify this general practice. These reasons are oriented toward either promoting testing efficiency or providing protection against illegitimate score gains. Each of these reasons will now be presented and discussed.

Item Dependence

Because proficiency tests contain multiple items, there is the possibility that the stem and/or the response options in one item may provide an examinee a clue regarding the answer to another item. Whenever an examinee recognizes such a clue, and thereby passes an item that he/she otherwise would have failed, an illegitimate score gain occurs. Denying item review is considered a means of protection against such score gains. Green (1988) commented on this issue:

> changing an answer to an earlier item probably promotes dependence of error components. One item may reveal the answer to another, or may remind the examinee of the method for solving an earlier problem.

Consequently, it would be better psychometrically to prohibit review. (p. 79)

It should be noted, however, that denying item review prevents only some of the potential score gain due to item clues. Even without item review, an item administered at a given point in a test can potentially provide a clue to any item administered after it. The problem of one item containing a clue to another item is therefore difficult to fully control during the test administration. An alternative solution to this problem is to study the items for clues and then constrain the item pool appropriately; for example, if item X is administered to a given examinee, and it provides a clue to item Y, then item Y cannot also be administered to that examinee. These types of item administration constraints are currently being used with a number of operational CATs, including the CAT version of the Graduate Record Examination.

Decreased Testing Efficiency

Another problem posed by item review is that it compromises the efficiency of the CAT. At each step in a typical adaptive testing procedure, the item from the pool is selected that provides maximum information (in an IRT sense) given the current estimate of the examinee's proficiency. This essentially means that item difficulty is matched to the examinee's current proficiency estimate. As the examinee's proficiency estimates converge during the test, so do the difficulty levels of the items administered. Once the adaptive portion of the test is completed, if the examinee is allowed to review his/her answers, any answer change that results in a previously failed item being passed or a previously passed item being failed will alter the examinee's sequence of proficiency estimates. Because the examinee's proficiency estimates have changed, the items that were administered are likely to be less well matched to the new sequence of proficiency estimates. This implies less efficient testing, as

the standard error of the final proficiency estimate is likely to increase as the result of review.

Probably the strongest statement against item review in relation to efficiency was provided by Wainer (1993), who asserted that "allowing an examinee to modify his/her responses will probably not yield a bias but will affect the efficiency of the test. And, since increased efficiency is the *raison d'etre* for CAT, we ought to avoid it" (p. 18). In Wainer's view, test features that detract from the efficiency of a CAT should be discouraged.

The limited research that has be conducted thus far on the effects of item review on CAT efficiency has not shown Wainer's (1993) concerns to be warranted. Three studies have been conducted in which the efficiency of a CAT both with and without item review has been compared. Lunz, Bergstrom, and Wright (1992), using a certification examination, found the decrease in test efficiency after review to be very small; the ratio of test efficiency was calculated to be .99. This indicated that, on the average, less than one additional test item would have been needed to recover the information lost due to changed answers. Similarly, Stone and Lunz (1994) found test efficiency ratios of .99 and .98 for two criterion-referenced tests. Lunz and Bergstrom (in press) found that item review had virtually no effect on the standard errors of the final proficiency estimates. Hence, the effect of decreased test efficiency due to the provision of item review appears to be negligible.

Vulnerability to Test Taking Strategies

Some researchers have expressed concerns that disingenuous examinees would be able use item review to artificially inflate their test scores through use of a deceptive test-taking strategy. Wainer (1993) described such a strategy in the context of a hypothetical situation in which examinees could both omit items and review answered items within a CAT. Although a scenario in which examinees

would be allowed to omit items within a CAT is unlikely to occur in practice, the strategy could be restated in a way that requires only item review. In this strategy, an examinee would intentionally choose incorrect answers for items, which would result in the items being administered becoming progessively easier. After completion of the adaptive portion of this relatively easy CAT, the examinee would then review his/her items and replace the intentionally wrong answers with as many correct answers as he/she could identify.

There are several noteworthy aspects to what I will term the "Wainer strategy." First, although it is clearly deceptive, successful use of the strategy would yield changed answer score gains that are legitimate, because the examinee's final answers to the administered items would reflect his/her state of knowledge regarding those items. Second, it requires the examinee to be proficient enough to identify the correct answers that are to be avoided prior to review and then chosen during review. Without such a level of proficiency, an examinee runs the risk of failing one or more relatively easy items and possibly ending up with a lower final proficiency estimate than would have been observed without the strategy. Third, the strategy represents an attempt to "beat" the invariance principle of IRT, which states that it should not matter whether an easier test was administered—the expected proficiency estimate should be the same as it would be with a more difficult test. The logic underlying the Wainer strategy is basically that if the test is easy enough, then invariance will not hold, and proficiency estimates will consequently be higher.

Two simulation studies have investigated the consequences of examinees attempting to use the Wainer strategy. Wang and Wingersky (1992) simulated a test-taking strategy in which examinees from a variety of proficiency levels intentionally provided the wrong answers to the first five items of a CAT and then tried to correctly answer the remainder of their items. After a

predetermined number of items had been administered, a response pattern was then generated that simulated the examinees' test performances if they had subsequently tried to answer all of their items correctly during review. The results indicated that, for low and moderate proficiency levels, mean post-review proficiency estimates were equal to the true proficiency values. For high proficiency levels, many high proficiency examinees were able to attain perfect test scores (which would yield unbounded maximum likelihood proficiency estimates), and thus were able to artificially boost their proficiency estimates through use of the Wainer strategy. For those examinees who did not attain pefect scores, however, the mean post-review proficiency estimates were markedly *lower* than the true proficiency values. Overall, Wang and Wingersky's results suggest that the Wainer strategy impacts test performance only for highly proficiency examinees; for those examinees, however, use of the strategy involves some risk.

Similar findings were recently reported by Gershon and Bergstrom (1995) who investigated the effects of simulating the Wainer strategy with a CAT whose item bank calibrations and scoring were based on the Rasch model. They also found that (a) examinees employing the strategy would be unlikely to increase their proficiency estimates and (b) the proficiency estimates of highly able examinees would tend to be negatively biased. Gershon and Bergstrom discussed the risk involved with employing the Wainer strategy:

> During review, in order to successfully cheat, the examinee must correctly answer *all* of the items that he purposefully missed. If an examinee makes a mistake, and fails to change even 1 answer from wrong to right, the consequences may be dire. (p. 6)

The results from Wang and Wingersky (1992) and Gershon and Bergstrom (1995) suggest that the Wainer strategy would not be overly attractive for

examinees to attempt. Examinees of low and moderate proficiency would not be advantaged; an examinee's expected proficency estimate would equal his/her true proficiency. Although examinees of high proficiency might pass all of their items, the strategy may place a great deal of stress on an examinee because failure on even a single easy item may lead to an underestimate of true proficiency. Moreover, even passing all of the items may lead to an underestimate of true proficiency if the Rasch model is used in scoring. It should also be noted that examinees of high proficiency are, in most testing contexts, those least in need of a deceptive test taking strategy. Nevertheless, the reported studies are simulations, and it remains unclear how successful actual examinees would be in attempting the Wainer strategy.

Even if the Wainer strategy could be successfully used, it should be relatively easy to spot examinees using such a strategy, and to take corrective action. For instance, the examinees could be told that if the difference between their proficiency estimates before and after review exceeded some threshold, then their test scores could be deemed invalid or additional test items would need to be administered.

A second strategy is based on a characteristic of a CAT that was first noted by Green, Bock, Humphreys, Linn, and Reckase (1984), who discouraged the provision of item review, noting that

> to the extent that the applicant knows or concludes that item difficulty depends on previous responses, the perceived difficulty of the present item may be taken as a clue to the correctness of the earlier responses. It *is* a form of feedback, however subtle; thus, permitting retracing would seem improper. (p. 356)

Kingsbury (personal communication, July 20, 1995) described how this information could be used by examinees in attempting to illegitimately increase

their scores. After answering a given item, if the examinee could discern that the succeeding item was more difficult, then he/she would know that the answer to the previous item was correct. If the succeeding item was less difficult, then the answer to the previous item was incorrect, and the examinee would know to change his/her answer to the previous item during review. It is not clear, of course, how accurately examinees would actually be able to discern whether the difficulty levels of successive items increased or decreased, particularly later in the test, when the changes in difficulty between successive items become relatively small. If, however, examinees are able to successfully use the Kingsbury strategy, then it would likely yield higher score gains than a successfully applied Wainer strategy, because a more difficult set of items would be administered with the Kingsbury strategy. Moreover, it would be very difficult for test givers to spot the Kingsbury strategy being used.

It should be noted that, although few examinees would, on their own, think to use either of these strategies, many could be coached to do so. It is certainly possible that unscrupulous test preparation companies would provide such coaching under the guise of "test taking" skills. This is potentially a very serious problem, and research efforts should be directed toward understanding the vulnerability of CAT to test-taking strategies designed to yield deceptive or illegitimate score gains.

Increased Testing Time

A consequence of providing item review is that it requires additional testing time. The only study reporting specific information on this issue was Vispoel et al. (1992), who found that providing review on computer-based tests increased average testing time by about 27%. This represents a distinct aspect of testing efficiency, because increased testing time implies that an examinee's proficiency is being measured less efficiently.

Because of the increased testing time required, the provision of item review on a timed test would need to be accompanied by a corresponding increase in the time limit imposed. Otherwise, item review would realistically be available only to those examinees who quickly completed the adaptive portion of the test, and thus had time remaining for review.

More Complicated Test Development

Providing item review certainly renders the task of CAT software development more complicated, partly due to the fact that examinees review items in different ways. Some examinees are motivated to review each answer, while others wish to review only a small number of items. Still others prefer, when first considering their test items, to mark some to be reviewed later. Conceptually, it is challenging to devise a system that is flexible enough to meet the diverse reviewing styles of examinees without being difficult to understand and/or awkward for examinees to use. If the review system is not easy to use, then examinees who would otherwise choose review might instead decline to use it. The sheer challenge of devising a functional review system has probably discouraged its development and use.

Arguments In Favor of Providing Review

In recent years, researchers and practitioners have begun to question the view that item review is inappropriate in a CAT. An argument for providing review has emerged that focuses largely on facilitating legitimate score gains, as opposed to protecting against illegitimate gains. The two primary reasons in favor of providing review will now be addressed.

Examinee Preference for Review

As group-administered multiple-choice tests came into widespread use, beginning with the U.S. Army Alpha test during World War I, examinees were allowed to review their test items. This does not appear to be a planned feature

of paper-and-pencil tests; rather, item review is an essentially uncontrollable examinee behavior that has come to be viewed by examinees as a normal aspect of taking a test. This perception has generalized to computer-based testing. Based on attitudinal data collected after the administration of a CAT without item review, both Baghi, Ferrara, and Gabrys (1992) and Legg and Buhr (1992) found that examinees reported being bothered by not being allowed to review items. Similarly, other studies (Vispoel, Rocklin, & Wang (1994); Vispoel et al., 1992) have found that examinees strongly favor the inclusion of item review in computer-based tests. Hence, there are clear indications that examinees desire item review with CATs.

Review Yields Legitimate Score Gains

A major component of the argument for providing item review is the sizable body of research on answer changing that has consistently found score gains. As indicated by the Waddell and Blankenship (1995) meta-analysis, most examinees will change an answer to at least one item on a test, those who change answers are likely to increase their scores, and score gains from answer changing average about 3%. This implies that denying item review will tend to lower examinee scores.

A handful of studies have investigated the effects of answer changing in the context of a CAT. In each of these studies (Lunz et al., 1992; Lunz & Bergstrom, 1994; Lunz & Bergstrom, in press; Stone & Lunz, 1994; Vispoel et al., 1992) the mean proficiency estimate was higher after item review. It therefore appears that the effects of answer changing within a CAT do not differ markedly from those with a paper-and-pencil test.

The finding of score gains on a CAT as the result of item review raises the question of whether the gains observed were legitimate or illegitimate. Although there does not appear to be any research regarding reasons for answer changing

with a CAT, the research with paper-and-pencil tests is consistent and probably generalizes well to a CAT context.

McMorris and Weideman (1986) first studied why examinees changed answers, reporting the following five reasons, ordered from most to least popular: rethinking the item and conceptualizing a new answer (57%), rereading the item and understanding the question better (28%), making a clerical error (8%), finding a clue in an item (3%), and finding a clue in another item (3%). The three most popular reasons should be considered legitimate reasons, because they produced a better representation of an examinee's state of knowledge at the beginning of the test. The two least popular reasons should be considered illegitimate, because they misrepresented the examinee's proficiency whenever they led to correct answers. Hence, in the McMorris and Weideman study, only 6% of the answer changes were reported to be for reasons that would lead to illegitimate score gains.

Similar results were found in subsequent studies. McMorris, DeMers, and Schwarz (1987) found that only 3% of the answer changing reasons were due to clues within or across items, while for Schwarz, McMorris, & DeMers (1991) and McMorris et al. (1991), the comparable values were 14% and 17%, respectively. Shatz and Best (1987) reported that 15% of the answer changes in their study were attributed to clues across items. Ramsey, Ramsey, and Barnes (1987) reported that 89% of the answer changes were for reasons that should be considered legitimate.

Taken together, the research on answer changing reasons suggests that examinees change answers for legitimate reasons about 90% of the time, and for illegitimate reasons about 10% of the time. This implies that the score gains that have been observed in studies of answer changing within a CAT are overwhelmingly due to legitimate reasons.

Evaluating the Arguments

A consideration of the arguments for and against item review is complicated by the presence of multiple stakeholders in the measurement process. One stakeholder is the test giver—the developer(s) of the test items and the accompanying test administration procedures. The primary goal of the test giver is to accurately assess an examinee's level of proficiency. But because it is difficult to assess with confidence how well this goal is being attained, test givers often rely on a strategy of developing an item pool that is representative of the domain of interest and then striving to produce reliable test scores using that item pool. In addition, test givers are motivated by time issues. In many testing programs, a large number of examinees must be tested in a limited amount of time. The capability to provide relatively rapid testing is one of the most attractive features of CAT to test givers. Because item review would require additional testing time, and would diminish the time savings offered by a CAT, test givers may be reluctant to decide to provide review.

A second stakeholder is the test taker—the examinee, whose goals and motivations differ, in some ways, from those of the test giver. Many examinees are interested in gaining accurate assessments of their proficiencies, which is a goal they share with the test giver. However, examinees are also motivated to simply score as well as possible, even if their scores are augmented by illegitimate gains. High test performance brings with it many societal rewards, such as a good course grades, college entrance, employment, promotion, or professional licensure. It is certainly no revelation that many examinees view tests more as obstacles to success than as valued sources of information regarding proficiency levels. The competitive atmosphere that often surrounds testing tacitly encourages examinees to seek high scores, even through illegitimate score gains. Conversely, test givers are encouraged to adopt

practices that will thwart testing strategies that could lead to illegitimate score gains.

Each stakeholder has a distinct perspective on the measurement process, and both perspectives are important to consider when designing a test. It is important to note that the arguments against item review are primarily in the interests of the test giver, while the arguments in favor of item review are primarily in the interests of the examinee. It is therefore difficult to weigh the arguments without taking into account both stakeholders' perspectives. The decision regarding whether or not to provide item review ultimately reduces to a matter of whose interests carry a greater weight. And because the decision makers are typically the test givers, it is perhaps not surprising that item review has been rarely provided within CATs.

Implicit in a decision not to provide item review is the assumption that a single exposure to a test item is sufficient. It certainly is sufficient from the standpoint of the CAT algorithm; the test cannot proceed until a response is given to an item, and once that response is given, the algorithm can update the proficiency estimate and choose the next item to administer. And, clearly, the CAT will proceed with the greatest efficiency if only a single item exposure is needed. If one assumes that examinees can either ascertain the correct answer to an item or cannot, and those who can will supply the correct answer on request, why should more than one exposure be needed? This is a mechanistic view of an examinee's test performance—analogous to that of a computer program checking a storage location in memory. Is the desired information stored in memory? If so, then retrieve it. If not, then there is no reason to check again.

This mechanistic model, however, is a poor representation of the examinee testing experience. Moreover, it is a subtle version of the enduring "your first answer is best" myth that persists regarding the advisability of answer changing.

People commonly experience instances in which they provide an answer to a problem, and then upon further consideration, realize that the correct answer was not the one they initially gave. It is also likely that insight regarding an item's correct answer can emerge later in the testing session without the examinee consciously thinking about the item. On an intuitive level, then, the mechanistic model does not match our test taking experiences.

The inadequacy of the mechanistic model is further revealed by the research literature on answer changing. When examinees re-consider their items, they have consistently been found to make changes that tend to improve their scores. Moreover, the score gains are largely legitimate, as indicated by the most common reasons given by examinees for their answer changes: "rethinking the item and conceptualizing a new answer" and "rereading the item and understanding the question better." These reasons are consistent with our current understanding of human reading, memory, and problem solving.

When we use a computer to administer a test, we must be careful not to adopt the expectation that examinees should behave in a computer-like fashion. Norman (1993) provided a salient commentary on this expectation:

> Society has unwittingly fallen into a machine-oriented orientation to life, one that emphasizes the need of technology over those of people, thereby forcing people into a supporting role, one for which we are most unsuited. Worse, the machine-centered viewpoint compares people to machines and finds us wanting, incapable of precise, repetitive, accurate actions. Although this is a natural comparison, one that pervades society, it is also a most inappropriate view of people. (p. xi)

Important to a discussion of item review is the issue of test score validity. When item review is allowed, are the resulting scores more or less valid indicators of proficiency than when review is denied? Based on the arguments

presented and the accompanying empirical evidence, it appears that providing item review on a CAT will yield more valid scores. If answer changing tends to yield score gains, and and these gains are largely legitimate, then it follows that the net result of item review will be more valid scores.

It might be argued that (a) the goal of measurement is often the identification of relative levels of proficiency (i.e, norm-referenced measurement) and not necessarily absolute levels, and (b) if item review is denied for all examinees, then it represents an equitable testing practice. This assertion requires the assumption that denying item review has an equal impact on the test performances of all examinees. Because there is little empirical evidence to support or refute this assumption, it is difficult to assess. There is evidence, however, that examinees vary substantially in the numbers of answers that they change (e.g., Lunz & Bergstrom, 1994) and that Black examinees make more answer changes than White examinees (Payne, 1984). It is possible that there are individual difference variables that moderate answer changing behavior, which in turn may moderate the impact of denying item review within a CAT.

Conclusions

The question of whether or not to provide item review within a CAT is one without a clear answer. Examinees express a desire for item review, and providing the opportunity for answer changing tends to result in legitimate score gains. On the other hand, item review requires additional testing time, and examinees may be able to improve their scores through clever test-taking strategies. This potential vulnerability to illegitimate score gains must be satisfactorily addressed in order to maintain the integrity of the test and the validity of the test scores. It is not acceptable, however, for the interests of the examinee to be sacrificed in the process. Testing practices that do not allow examinees to exhibit their true levels of proficiency are fundamentally flawed. In

our pursuit of reliable, efficient testing, this point can be overlooked. As Wiggins (1993) warned, "tests are intrinsically prone to sacrifice validity to achieve reliability and to sacrifice the student's interests for the test makers" (p. 4).

Computer-based testing practice should be responsive to both the interests of the test giver and those of the examinee. If these often-conflicting interests cannot be reconciled, practitioners should seek alternative testing methods. In the case of item review, the problems posed both by providing and denying item review may encourage a search for innovative types of computer-based tests that are efficient, yet promote optimal, legitimate examinee performance.

References

Baghi, H., Ferrara, S. F., & Gabrys, R. (1992, April). *Student attitudes toward computer-adaptive test administrations*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Benjamin, L. T., Cavell, T. A., & Shallenberger III, W. R. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11,* 133-141.

Gershon, R., & Bergstrom, B. (1992, April). *Does cheating on CAT pay: Not!* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Green, B. F. (1988). Construct validity of computer-based tests. In H. Wainer & H. I. Braun (Eds.), *Test validity* . Hillsdale, NJ: Lawrence Erlbaum Associates.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347-360.

Legg, S. M., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice, 11,* 23-27.

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement, 31,* 251-263.

Lunz, M. E., & Bergstrom, B. (in press). Computerized adaptive testing: Tracking candidate response patterns. *Journal of Educational Computing Research*.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement, 16,* 33-40.

McMorris, R. F., DeMers, L. P., & Schwarz, S. P. (1987). Attitudes, behaviors, and reasons for changing responses following answer-changing instruction. *Journal of Educational Measurement, 24,* 131-143.

McMorris, R. F., Schwarz, S. P., Richichi, M. F., Buczek, N. M., Chevalier, L. C., & Meland, K. A. (1991). *Why do students change answers on tests?* (Report No. TM 017 971). Albany, NY: State University of New York. (ERIC Document Reproduction Service No. ED 342 803)

Norman, D. A. (1993). *Things that make us smart*. Reading, MA: Addison-Wesley.

Payne, B. D. (1984). The relationship of test anxiety and answer-changing behavior: An analysis by race and sex. *Measurement and Evaluation in Guidance, 16,* 205-210.

Ramsey, P. H., Ramsey, P. P.. & Barnes, M. J. (1987). Effects of student confidence and item difficulty on test score gains due to answer changing. *Teaching of Psychology, 14,* 206-210.

Schwarz, S. P., McMorris, R. F., & DeMers, L. P. (1991). Reasons for changing answers: An evaluation using personal interviews. *Journal of Educational Measurement, 28,* 163-171.

Shatz, M. A., & Best, J. B. (1987). Students' reasons for changing answers on objective tests. *Teaching of Psychology, 14,* 241-242.

Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurment in Education, 7,* 211-222.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 7,* 53-79.

Vispoel, W. P., Wang, T., de la Torre, R., Bleiler, T., & Dings, J. (1992, April). *How review options and administration modes influence scores on computerized vocabulary tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Waddell, D. L., & Blankenship, J. C. (1994). Answer changing: A meta-analysis of the prevalence and patterns. *The Journal of Continuing Education in Nursing, 25,* 155-158.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12,* 15-20.

Wang, M. & Wingersky, M. (1992, April,). *Incorporating post-administration item response revision into a CAT.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wiggins, G. P. (1993). *Assessing student performance.* San Francisco: Jossey-Bass.

Footnotes

[1]In this discussion, it is assumed that dichotomously-scored multiple-choice items are administered, which the most commonly used item type in a CAT. Most, if not all, of the ideas expressed here, however, generalize to CATs administering polytomous and/or constructed response item types.

Table 1

Selected Results from Two Review Articles on Answer Changing

Research Finding	Benjamin et al. (1984)	Waddell & Blankenship (1995)
Outcome of Answer Changes		
Incorrect to Correct	58%	57%
Incorrect to Incorrect	23%	22%
Correct to Incorrect	20%	21%
Effect of Answer Changes on Final Scores		
Higher Score	68%	68%
Same Score	14%	17%
Lower Score	15%	15%